ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Ad-Hoc Method for Social Data Collection Using Foursquare

Shameena V [1], Sentilkumar [2]

P.G. Student, Department of Computer Science and Engineering, SVHEC Engineering College, Gopichettipalayam, Tamilnadu, India[1]

Associate Professor, Department of Computer Science and Engineering, SVHEC Engineering College, Gopichettipalayam, Tamilnadu, India[2]

ABSTRACT: Data generated on location-based social networks provide rich information on the whereabouts of urban dwellers. Specifically, such data reveal who spends time where, when, and on what type of activity (e.g., shopping at a mall, or dining at a restaurant). That information can, in turn, be used to describe city regions in terms of activity that takes place therein. For example, the data might reveal that citizens visit one region mainly for shopping in the morning, while another for dining in the evening. Furthermore, once such a description is available, one can ask more elaborate questions. For example, one might ask what features distinguish one region from another—some regions might be different in terms of the type of venues they host and others in terms of the visitors they attract. As another example, one might ask which regions are similar across cities. In this paper, we present a method to answer such questions using publicly shared Foursquare data. Our analysis makes use of a probabilistic model, the features of which include the exact location of activity, the users who participate in the activity, as well as the time of the day and day of week the activity takes place. Compared to previous approaches to similar tasks, our probabilistic modeling approach allows us to make minimal assumptions about the data—which relieves us from having to set arbitrary parameters in our analysis (e.g., regarding the granularity of discovered regions or the importance of different features). We demonstrate how the model learned with our method can be used to identify the most likely and distinctive features of a geographical area, quantify the importance features used in the model, and discover similar regions across different cities. Finally, we perform an empirical comparison with previous work and discuss insights obtained through our findings.

KEYWORDS: Urban computing, location based social network, cities, user activity

I. INTRODUCTION

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered.

Web Server Data: The user logs are collected by the Web server. **Typical** data includes IP address, page reference and access time.
Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events.
Recommender System: Recommender systems or recommendation systems are a subclass of information filtering system that seek to predict the "rating" or "preference" that a user would give to an item. Recommender systems

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

typically produce a list of recommendations in one of two ways – through collaborative and content-based filtering or the personality-based approach. Analyzing a large amount of information on users‟ behaviors, activities or preferences and predicting what users will like based on their similarity to other users. Collaborative filtering approaches building a model from a user's past behavior as well as similar decisions made by other users. This model is then used to predict items (or ratings for items) that the user may have an interest in. A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore it is capable of accurately recommending complex items such as movies without requiring an "understanding" of the item itself.

II.RELATED WORK

Mobile networks enable users to post on social media services (e.g., Twitter) from anywhere. The activities of mobile users involve three major entities: user, post, and location. The interaction of these entities is the key to answer questions such as who will post a message where and on what topic? In this paper, we address the problem of profiling mobile users by modeling their activities, i.e., we explore topic modeling considering the spatial and textual aspects of user posts, and predict future user locations. We propose the first ST (Spatial Topic) model to capture the correlation between users' movements and between user interests and the function of locations. We employ the sparse coding technique which greatly speeds up the learning process. We perform experiments on two real life data sets from Twitter and Yelp. Through comprehensive experiments. Spatial item recommendation has become an important means to help people discover interesting locations, especially when people pay a visit to unfamiliar regions. Some current researches are focusing on modeling individual and collective geographical preferences for spatial item recommendation based on users' check-in records, but they fail to explore the phenomenon of user interest drift across geographical regions, i.e., users would show different interests when they travel to different regions. Besides, they ignore the influence of public comments for subsequent users' check-in behaviors. Specifically, it is intuitive that users would refuse to check in to a spatial item whose historical reviews seem negative overall, even though it might fit their interests. Therefore, it is necessary to recommend the right item to the right user at the right location. In this paper, we propose a latent probabilistic generative model called LSARS to mimic the decision-making process of users' check-in activities both in home-town and out-of-town scenarios by adapting to user interest drift and crowd sentiments, which can learn location-aware and sentiment-aware individual interests from the contents of spatial items and user reviews. Due to the sparsity of user activities in out-of-town regions, LSARS is further designed to incorporate the public preferences learned from local users' check-in behaviors. Finally, we deploy LSARS into two practical application scenes: spatial item recommendation and target user discovery. Extensive experiments on two large-scale location-based social networks (LBSNs) datasets show that LSARS achieves better performance than existing state-of-the-art methods.

III. PROPOSED ALGORITHM

Our work belongs to the growing field of Urban Computing and shares its motivation. First, as an ever increasing number of people live in cities, understanding how cities are structured is becoming more crucial. Such structure indeed affects the quality of life for citizens (e.g. how much time we spend commuting), influences real-life decisions (e.g., where to rent an apartment or how much to price a house), and might reflect or even enforce social patterns (e.g. segregation of citizens in different regions). Second, switching perspective from the city to the people, the increasing amount of data produced by urban dwellers offer new opportunities towards understanding how citizens experience their cities.

This understanding opens possibilities to improve the citizens' enjoyment of cities. For instance, by matching similar regions across cities,To improve the relevance of out-of-town recommendations for travelers. The data we use were generated on Foursquare, a popular location-based social network, and provide rich information about the offline activity of users. Specifically, one of the main functionalities of the platform is enabling its users to generate check-ins that inform their friends of their whereabouts. Each check-in contains information that reveals who (which user) spends time where (at what location), when (what time of day, what day of week), and doing what (according to the kind of venue: shopping at a grocery store, dining at a restaurant, and so on).

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

Best First Search for Joint Sims:
our goal is to uncover associations between geographic locations and other features of venues. Such associations are captured as k topics in the model each data point is assigned probabilistically to one topic and different topics generate data venues with different distributions of features. As an example, one topic might generate venues (data points) that are located in the south of a city (feature: location) and are particularly popular in the morning (feature: time of the day), while another might generate venues that are located in the north of a city (feature: location) and predominantly restaurants, bars, and night-clubs (feature: category).
Specifically, to generate one data point, the model per- forms the following steps:

• Select one (1) out of k available topics {1, 2, . . . , k} according to a multinomial probability distribution θ = (θ1 , θ2 , . . . , θk). Let the selected topic be z.

• Generate a geographic location loc = (x, y) from a bivariate Gaussian distribution with center c = cz and variance matrix Σ = Σz.

For the i-th categorical feature, generate a list u = ui of N = Ni items, where Ni is specified as input for this data point. Each element in the list is selected randomly with replacement from a set U = Ui = {u1 , u2 , . . . , um } according to multinomial probability β = βz = (βz |1 , βz |2 , . . . , βz |m), with βz |j ≥ 0 and i j=1..m βz |j = 1.21

To the best of our knowledge, the problem of identifying similar regions across cities has not been defined formally before within a probabilistic framework. In this propose the first algorithm, GeoExplore, Algorithm GeoExplore follows a typical best-first exploration scheme and comprises of the following two phases.

1. Its first phase consists of one step: it begins with a candidate collection of regions G1 and G2 for each side (let us call them ‘base regions’) and evaluates all pair wise similarities jointsim (G1, G2), for G1 ∈ G1, G2 ∈ G2.
2. Its second phase consists of the remaining steps: it explores the possibility to improve the currently best jointsim measure by combining previously considered regions. This is motivated by the fact that Problem 4 favors regions of larger probability mass, and therefore combined regions might yield better jointsim values.

IV. PSEUDO CODE

Input: models I1 and I2 , base regions G1 , G2

Output: Best Pair G1 , G2

 # INITIALIZE

BestG1 = NULL, BestG2 = NULL, BestScore = 0

H = MaxHeap()

 # Initialize max-heap with empty solution, zero score

Push(BestG1 , BestG2 , BestScore) to H

while H is Not Empty do

 # RETRIEVE top solution in max-heap

Pop (G1 , G2 , Score) from H

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

 # UPDATE best solution

if Score > BestScore then

BestG1 = G1 , BestG2 = G2 , BestScore = Score

end if

 # EXPAND retrieved solution

for Ga , Gb in expand(G1), expand(G2) = G1 , G2 do

Score = jointsim(Ga , Gb |I1 , I2)

Push (Ga , Gb , Score) to H

end for

end while

return best pair

It repeats a three-steps Retrieve - Update - Expand procedure for each step. During Retrieve, the algorithm retrieves the next candidate solution. Each candidate solution comes in the form of a triplet; two Gaussians and their jointsim score. During Update, the algorithm updates the score of the best-matching pair, if a better pair has just been retrieved. Finally, during Expand, the algorithm expands the latest retrieved Gaussians to form Gaussians from each side, and thus new candidate solutions or Subroutine expand(G, I) operates as follows:

• When G is not specified (i.e., G = NULL in Algorithm 1), then expand simply returns the set of base regions G. This case occurs during the first expansions only. Moreover, each base region Gi ∈ G is associated with positive weight wi, either specified as input, or set to 1/kGk by default.

• When G = Gi for some Gi ∈ G, then expand returns the set of Gaussians {Gi} S {Gi ∪ Gj ; Gj ∈ G, j 6= i}, where Gi ∪ Gj is defined as the best Gaussian-fit to the mixture model determined by [Gi , Gj], with respective proportions (wi, wj). The intuition for this step is that we expand the best-performing pair of Gaussians by combining them with other base Gaussians.

• In a recursive fashion, when G = Gi ∪ Gi 0 . . . Gi 00 , then expand returns the set of Gaussians {G} S {G∪Gj ; Gj ∈ G, j 6= i, i0 , . . . , i00} each defined as the best Gaussian- fit to the mixture model determined by [G, Gj], with respective proportions (wi + wi 0 + . . . + wi 00 , wj).to prevent the algorithm from exploring the combinatorially large space to terminate GeoExplore after a number R of while loops.

V.SIMULATION RESULTS

The url that are to be recommended will be identified based on ranking and similarity measure. The similarity measure is calculated among the users by comparing their similar interests. The users are grouped in certain cluster those having the similar preferences. The url that are grouped among the cluster will be sorted in the order. The url is sorted based on the user interests which having the highest priority. The data sparse problem can be removed. Hence, the web page that is to be recommended to the user is sorted where the recommendation will achieve the higher accuracy. The user unobserved preferences are also considered for the better recommendation. The problem of new user in the internet can also be solved using this method. Hence, Naïve-Bayesian algorithm is effective algorithm to recommend the web pages to the user.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

The geographic component of the model use does not match the notion of neighborhood as conceived in a more administrative sense, e.g., as a set of roads or other boundaries that enclose a geographical area. our goal in the paper is not to discover such neighborhoods, but rather to discover geographical patterns that represent well, in a probabilistic sense, the activity observed in the data at hand. The geographical patterns look for are associated with more loosely-defined regions, represented by two-dimensional Gaussians. Moreover, note that the model allows regions to overlap and at the same time accommodate a number of categorical features. In practice, this means that we might discover topics in the data that have highly overlapping geographical regions, but are differentiated based on other features. For example, in training the model, we might discover that there are two topics in the same region – one consisting of venues that operate early in the morning, and of venues that operate late in the evening. This is a basic difference from related work such that aims to directly partition the area of a city into non-overlapping regions

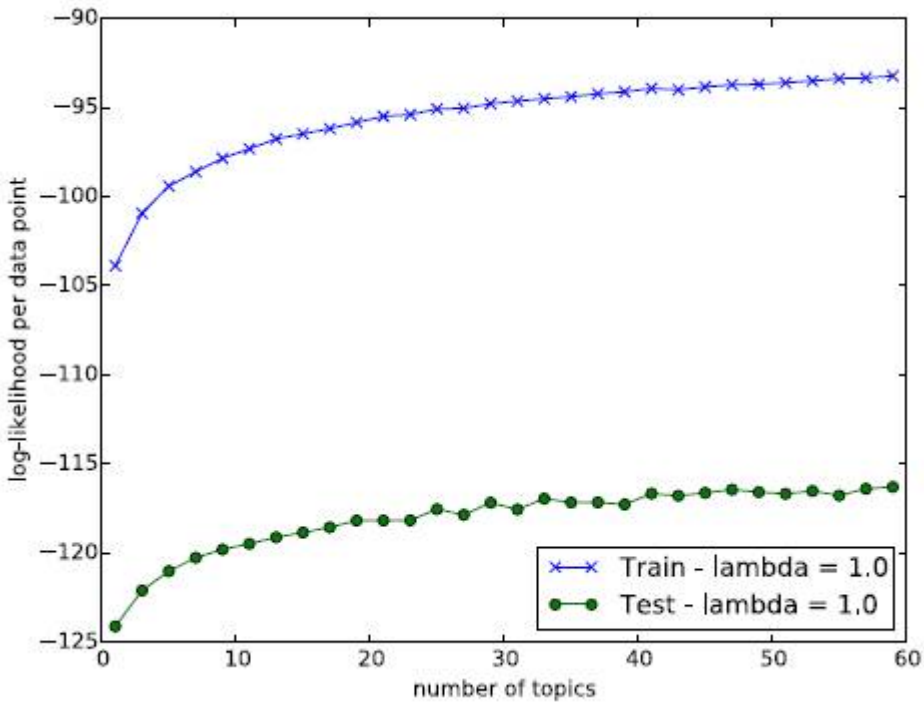


Fig.1. Log-likelihood per data point on the training and test datasets of Paris, for _ ¼ 1 and increasing k.

A max-likelihood vector m is computed once for each feature from the raw relative frequencies of observed values of that feature in the dataset. For fixed k and _, the maximum likelihood value of the remaining parameters can be computed with a standard expectation-maximization algorithm. The steps of the algorithms are provided below

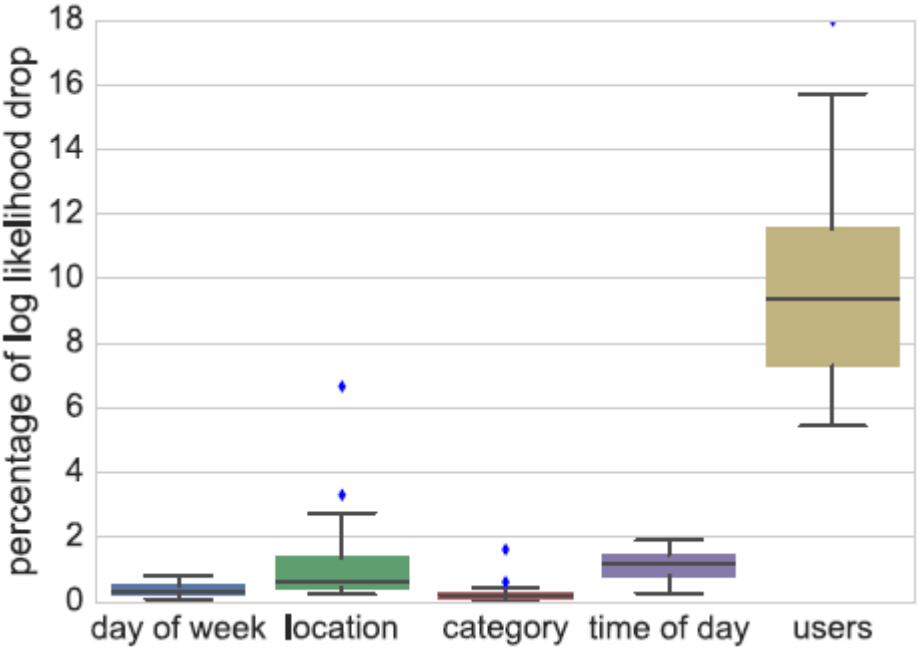
ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017



Contribution of each feature to the data likelihood. The box plots summarize how much the log likelihood drops once we fix the distribution of a single feature across topics. We observe a consistent behavior across cities, in that the variance of users across topics is most important for the predictive performance of the model

VI. CONCLUSION AND FUTURE WORK

The use of a probabilistic model to reveal that venues are distributed in cities in terms of several features. As most habitants of a city do not visit most of the available venues cope with the induced sparsity by adapting the sparse modeling approach of to data at hand. Fitting our model to a large dataset of more than 11 million checkins in 40 cities around the world First, using the extracted model instances is calculated the probability distribution of a single feature conditioned on the location in the city. This enabled us to construct a heat map of that feature to highlight what feature values are most likely and distinctive at different locations within a city

Secondly, it is described a principled approach to quantify comments. The importance of different features within the trained models. Whereas all features contribute and discovered that the most defining feature for the components uncovered by the model is the visitors of venues. This finding suggests that further analysis of user behavior is a promising direction for extracting additional insights.

Third, after focusing on the various regions of a single city, the extracted model instances to find the most two similar regions between two cities, a task which was previously attempted with a more heuristic approach .the solid theoretical grounds of probabilistic models to define a principled measure of similarity .

REFERENCES

1. Zibin Zheng, Hao Ma, Michael R. Lyu, Fellow, and Irwin King, (2011),"QoS-Aware Web Service Recommendation by Collaborative Filtering",IEEE transactions on
2. Freddy L´ecu´e,(2010)"Combining Collaborative Filtering and Semantic Content-based Approaches to Recommend WebServices", IEEE Fourth International Conference on Semantic Computing, pp. 200-205.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 11, November 2017

3. Alexandrin Popescul ,Lyle H. Ungar , David M. Pennock,Steve Lawrence, (2001), "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments", Published in Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pp. 437-444,August.
4. Byron Bezerra and Francisco de A. T. E Carvalho, (2004)," A Symbolic Hybrid Approach to Face the New User Problem in Recommender Systems", Springer-Verlag Berlin Heidelberg, pp. 1011–1016.
5. Katja Niemann and Martin Wolpers,(2015), "Creating Usage Context-Based Object Similarities to Boost Recommender Systems in Technology Enhanced Learning",IEEE transactions on learning technologies, Vol. 8, no. 3, pp. 274-285,September.
6. KazuyoshiYoshii,MasatakaGoto,KazunoriKomatani,TetsuyaOgata,HiroshiG.O
kuno, (2006),"Hybrid Collaborative and Content-based Music Recommendation Using Meghna Khatri, (2012), "A Survey of Naïve Bayesian Algorithms for Similarity in Recommendation Systems" , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, pp. 217-219, May.
7. Kebin Wang and Ying Tan, (2011),"A New Collaborative Filtering Recommendation Approach Based on Naïve Bayesian Method", Springer-Verlag Berlin Heidelberg, pp. 218–227.